DOCUMENT RESUME

ED 200 632 TM 810 218

AUTHOR Douglass, James B.
TITLE Item Bias, Test Speededness, and Rasch Tests of Fit.
SPONS AGENCY Educational Testing Service, Princeton, N.J.; Michigan State Univ., East Lansing.
PUB DATE Apr 81
NOTE 18p.; Paper presented at the Annual Meeting of the American Educational Research Association (65th, Los Angeles, CA, April 13-17, 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Black Students; *Computer Programs; Correlation; *Difficulty Level; *Goodness of Fit; Higher Education; Latent Trait Theory; Mathematical Models; Racial Bias; Response Style (Tests); Sex Differences; *Test Bias; *Test Items; Vocabulary; White Students
IDENTIFIERS BICAL Computer Program; Factor Invariance; Item Position (Tests); Kuder Richardson Reliability Formulas; LOGIST Computer Program; *Rasch Model; Spearman Rank Correlation Coefficient; *Speededness (Tests)

ABSTRACT

Relationships between item bias, item difficulty invariance, Rasch tests of item fit, and item position in a speeded 72-item Michigan State University Vocabulary Placement Test were investigated using 143 black males, 289 black females, 2,953 white males and 3,271 white females. Measures of item bias and item difficulty invariance were determined and compared for the one-parameter logistic item-response model using the BICAL and LOGIST computer programs. The emphasis was not upon detection of bias, but rather upon the effects of bias. The BICAL program scores as incorrect those items omitted or not reached by an examinee. The LOGIST program treats item responses differently depending upon whether an item is considered to be correctly answered, incorrectly answered, omitted or not reached. In this study, the BICAL estimates were found to be subject to ethnic bias due to test speededness; this effect was not present in the LOGIST estimates. Item difficulty invariance was related to several Rasch tests of fit for the BICAL estimates. (Author/RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED200632

U S DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ITEM BIAS, TEST SPEEDEDNESS, AND RASCH TESTS OF FIT

by

James B Douglass

Educational Testing Service

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. B. Douglass

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 810 218

Paper presented at the 1981 Annual Meeting of the American Educational Research Association, Los Angeles, CA, April 13-17, 1981. The author gratefully acknowledges the assistance received from Debbie Bennett and Dr. Charles Eberly and the review of an earlier draft by Linda Cook and Dan Eignor. Financial support for this research has been provided by ETS, the Office of Orientation Placement Testing, the Learning and Evaluation Service and the Provost's Office of Michigan State University.

ABSTRACT

Relationships between item bias, item difficulty invariance, Rasch tests of item fit, and item position in a speeded 72-item vocabulary placement test were investigated using 143 Black males, 289 Black females, 4,953 White males and 3,271 White females. Measures of item bias and item difficulty invariance were determined and compared for the one-parameter logistic item-response model using the BICAL and LOGIST computer programs. The BICAL estimates were found to be subject to ethnic bias due to test speededness; this effect was not present in the LOGIST estimates. Item difficulty invariance was related to several Rasch tests of fit for the BICAL estimates.

Item response theory (IRT), sometimes called latent trait or item characteristic curve theory, has been applied to a number of practical testing problems in recent years (e.g., Douglass, 1980; Lord, 1977; Marco, 1977; and Rentz and Bashaw, 1975). One promising application of IRT is to the problem of investigating item bias. Item response theory methods comprise but one of several approaches to the problem of empirically determining degrees of item bias (Rudner, Getson, and Knight, 1980). Definitions and empirical techniques of determining item bias can be visualized as existing on a continuum anchored at one end by bias in test use and at the other by bias that exists within a test. Item response theory methods, which assume that a single ability determines the probability of answering an item correctly, tend to fall toward the bias-within-test end of the continuum. In this context, item bias can be conceptualized as the extent to which an item measures different skills for different groups of people. Operationally, item bias can be characterized as the extent to which there are different probabilities of correct response to an item for examinees of the same ability level who have different group affiliations. It should be emphasized that the presence of item bias does not imply bias in test usage.

The present study focuses on the one-parameter logistic item-response model, sometimes called the Rasch model, and upon bias due to the speededness of a test. The primary purpose of the study is to compare the way in which item parameter estimates from two different operationalizations of the one-parameter model are affected by test speededness. The emphasis is not upon detection of bias, but rather upon the effects of bias. One operationalization depends upon item parameter estimates resulting from the use of the BICAL (Wright, Mead and Bell, 1979) computer program, where items omitted or not reached by an examinee are scored as incorrect. The second operationalization uses the same one-parameter model, but depends upon estimates from the LOGIST (Wood, Wingersky and Lord, 1976) computer program. This program treats item responses differently depending upon whether an item is considered to be correctly answered, incorrectly answered, omitted or not reached.

The primary objectives of this study are to compare, using item parameter estimates from the BICAL and LOGIST computer programs based on a speeded test given to Blacks, Whites, males and females:

1) Relationships between item bias (and invariance) estimates and item position within the test.

2) Relationships between item bias (and invariance) estimates and selected Rasch tests of fit.

3) Magnitudes of true scores across the first and second halves of the test for a simulated group of examinees.

Method

Sample

The examinees consist of 6,656 students planning to enter Michigan State University (M.S.U.) in the fall of 1979, who took the M.S.U. Vocabulary Placement Test during the summer of 1979. All entering freshmen at Michigan State are required to take a battery of placement examinations, including the vocabulary test. Of the 6,656 examinees, 143 are Black males (BM), 289 are Black females (BF), 2,953 are White males (WM), and 3,271 are White females (WF). The White females were divided into two groups. The first group (WF1), more typical of the entire sample of White females, consists of the 3,121 women with the highest student numbers. The second group (WF2), of higher average ability than the total sample of White females, includes the 150 White females, with the lowest student numbers. Student numbers are assigned to Michigan State applicants in the order in which the applicants are admitted to the university. Since the clearly acceptable students are admitted immediately upon receipt of their applications, those admitted first tend to be academically stronger than those admitted later.

The M.S.U. Vocabulary Test consists of 72 four-option multiple-choice items. Each item stem is a word given in the context of a word phrase. The options are words or word phrases. The correct option is the synonym of the word presented in the stem. Items are arranged alphabetically.

Measurement Model

The one-parameter logistic item-response model was the measurement model used for the study. The one-parameter model specifies a mathematical relationship between person ability and the probability of the correct response to an item that depends only upon one person parameter, called ability, and one item parameter, called difficulty. The one-parameter model defines the probability that a person, g, will answer an item, i, correctly as

$$P_{ig} = (1+e^{-K(\theta_g - b_i)})^{-1} \qquad (1)$$

where K is an arbitrary positive constant and e is the base of the natural logarithms, which is equal to 2.71828.... θ_g is the ability of person g and b_i is the difficulty of item i. Due to the mechanics of the BICAL computer program, K is equal to unity.

In order to provide a closer correspondence between the normal ogive and logistic models, in the LOGIST program K is equal to approximately 1.7 times an item discrimination value (Wood, Wingersky and Lord, 1976). By specifying the item discrimination to be 1/1.7 for all items, K will equal unity, which will insure that parameters of the same one-parameter model will be estimated by both programs.

Estimation of Model Parameters

Although both the LOGIST and BICAL computer programs can be used to estimate person ability and item difficulty parameters based on the one-parameter item-response model, there can be considerable differences in the resulting estimates. The differences are primarily caused by the fact that the programs maximize different criterion functions. The BICAL computer program computes estimates of item difficulties, b_i, (i=1, 2,..., n) and person abilities, θ_g, (g=1, 2,..., N), that maximize the likelihood function

$$L = \prod_{g=1}^{N} \prod_{i=1}^{n} P_{ig}^{u_{ig}} Q_{ig}^{1-u_{ig}} \qquad (2)$$

where n is the number of items and N is the number of examinees. Q_{ig} is equal to $(1 - P_{ig})$ and $u_{ig} = 1$ if examinee g answers item i correctly and $u_{ig} = 0$ otherwise.

Frequently in practical testing situations, examinees fail to answer items in a test. They may decide to skip over an item after reading it or they may not have enough time to reach items at the end of a test. For input to the LOGIST computer program, items consecutively omitted by an examinee at the end of an examination are usually coded "not reached" for that examinee. Items not answered but followed by one or more items that are answered correctly or incorrectly are coded as omitted. Although some items coded omitted may not have been read and some items coded not reached may have been read, but not answered, as Lord (1980, p. 182) points out, the assumption made in coding items omitted and not reached seems the best practical assumption currently available. Following this line of reasoning the likelihood function was modified for the LOGIST computer program (Lord, 1974). The modification results in the criterion function

$$L' = \prod_{g=1}^{N} \prod_{i=1}^{n_g} P_{ig}^{v_{ig}} Q_{ig}^{1-v_{ig}} \qquad (3)$$

where n_g is the number of items finished by examinee g. If examinee g answered item i correctly or incorrectly, $v_{ig} = u_{ig}$. If examinee g omitted item i, then $v_{ig} = 1/A_i$, where A_i is the number of options for item i. LOGIST estimates the person and item parameters that maximize the criterion function, L*.

It can be seen by examining the criterion function that items coded not reached for an examinee are not used in estimating his/her ability. It is also apparent that the difficulty of an item toward the end of a partially-speeded test is not over estimated by assuming that the response would have been incorrect if the item had been reached. The $v_{ig} = 1/A_i$ for an omitted item i, is based on the assumption that an examinee's chance of answering omitted items correctly, were he/she to answer the item, is approximately $1/A_i$, i.e., chance level. This should be a more reasonable assumption than the assumption that the omitted item would have been answered incorrectly. Of course as the likelihood function (2) used in the BICAL program implies, omitted and not reached items are treated as incorrect in BICAL.

Procedure

The BICAL and LOGIST computer programs were used to estimate item difficulties for the 72-item vocabulary test using each of the five subgroups of examinees previously described. This resulted in ten independent sets of item parameter estimates. For the one-parameter model, the scale underlying item difficulties and person abilities is unique only up to an additive constant. The average item difficulty is set to zero within the BICAL program for any set of items. Since the items were the same for each subgroup, no further scaling was necessary. The LOGIST option which bases scaling on item difficulty estimates was used to insure that all of the LOGIST item difficulty estimates were on a common scale. The item discriminations were set so that BICAL and LOGIST estimates were based on the same basic item response model (1). Therefore, differences between the results of the two programs are caused by the differences in the functions that are maximized, (2) and (3).

Indices of item bias and invariance were calculated for each of the six possible pairs of Black males (BM), Black females (BF), White males (WM) and the larger group of White females (WF1). Indices were also calculated for the two groups of White females compared to one another (WF1 and WF2). Separate indices of bias and invariance were calculated using item difficulties resulting from BICAL and LOGIST. Although a number of item bias indices based on the one-parameter model are possible (see e.g., Rudner, Getson and Knight, 1980), the simple difference between item difficulty estimates across examinee groups was used. Items are considered to demonstrate bias if this difference is large in either direction. The bias is against the group with the larger difficulty estimate.

Item difficulty invariance across groups was defined as the absolute difference between item difficulties across the examinee groups. If item parameters are independent of the examinee groups used, then this difference should be small. Lack of invariance would be evidenced by a large difference, regardless of direction, in item difficulty estimates across examinee groups.

Spearman rank-order correlation coefficients were calculated to measure relationships between bias (and invariance) and other variables, including item position in the test, total t-fit, between t-fit, the absolute value of the deviation from unity of the discrimination index and item difficulty standard error. See Wright, Mead and Bell (1979) or Wright and Stone (1980) for a discussion of the various BICAL-produced Rasch fit statistics.

To further illustrate the effects of test speededness and the associated bias resulting from a test that is differentially speeded for different subgroups, a simulation study was done. A sample of 500 person abilities was generated using a computer routine that returns pseudo-random numbers from a Gaussian normal distribution with a mean of zero and a variance of unity. Expected observed scores, i.e., true scores, corresponding to the simulated abilities were calculated. The true score calculations were based on the item difficulty estimates that were obtained for each of the five groups of real examinees. True-score distributions were generated using BICAL and LOGIST estimates. The true scores were generated separately for the first and second halves of the test. A true score, T_g, for an individual g and a set of items I, is defined by

$$T_g = \sum_{i \in I} P_{ig} \qquad (4)$$

Of course, for the first half of the test the summation is over items 1 to 36 and for the second half the summation is over items 37 to 72. Comparisons were made between true-score means across the half tests.

Results

Descriptive statistics for the five groups of examinees on the 72-item M.S.U. Vocabulary Placement Test are presented in Table 1. The test is differentially speeded for the various groups. The three criteria of speededness used in this study are the percentage of examinees completing the test, the percentage of examinees completing 75% of the test, and the number of items reached by 80%

Table 1

Descriptive Statistics for the 72-item
M.S.U. Vocabulary Placement Test

Statistic	Group				
	Black Males (BM)	Black Females (BF)	White Males (WM)	White Females (WF1)	White Females (WF2)
N	143	289	2953	3121	150
Mean	25.5	24.1	37.2	35.4	40.0
Standard Deviation	11.7	9.7	11.5	10.7	10.1
Percent Completing Test	30.8	39.8	66.2	67.8	82.7
Percent Completing 75% of test	68.5	77.5	93.9	94.4	96.7
Item Reached By 80% of Group	50	53	64	65	72
KR 20	.91	.86	.90	.88	.87

of the group of examinees. All three criteria indicate that the test is most speeded for Black males, second most speeded for Black females, next most speeded for White males, and least speeded for White females. Among White females the test is less speeded for the group that was chosen because of earlier admission to the university. By all three criteria, the differences between Blacks and Whites are more pronounced than are the differences between males and females. Notice that although the test is more speeded for the males, they tend to obtain a somewhat higher number-right score, an indication of a differential effect of test speed and power for males and females.

Spearman rank-order correlations between item position in the test and measures of item bias and invariance are presented in Table 2. Item bias and invariance measures were determined using estimates independently obtained from the BICAL and LOGIST computer programs. In each case the measure of bias is the item difficulty estimate obtained from the first group indicated minus the item difficulty estimate obtained from the second group. Therefore, a positive correlation indicates that item bias against the first group is positively correlated with the serial position of the item in the test.

An examination of Table 2 reveals that, for estimates based on the BICAL computer program, all comparisons between Blacks and Whites show a significant relationship between item bias against Blacks and item position in the test. Correlations range from .30, for the comparison of Black females to White males, to .56 for the comparison of Black males to White males. The comparisons between males and females of the same ethnicity are not significant, .03 and .23, at the $p < .02$ level. Notice that BICAL-based estimates do result in a significant (.37, $p < .002$) relationship between bias and item position for the comparison between the regular group of White females and the selected group of White females. Also note that the relationships between item bias, resulting from the use of the BICAL computer program, and item position are consistent with the information on test speededness for the various groups provided in Table 1.

The correlations between item bias resulting from the use of the LOGIST computer program and item position range from -.13 to .11. None of these correlations are significant at even the .05 level. Similarly, none of the BICAL- and LOGIST-based measures of item difficulty invariance, ranging from -.18 to .22, are significant at the .05 level. Given the information contained in Table 2, it is not surprising that item position is significantly correlated with BICAL item difficulty estimates averaged over the five groups (.39, $p < .001$), but not with LOGIST item difficulty estimates (.10).

Table 2

Spearman Rank-Order Correlations Between Item Position and Measures of Item Bias and Invariance

	Bias		Lack of Invariance	
Group Comparison	BICAL	LOGIST	BICAL	LOGIST
BM-BF	.23	-.13	-.18	-.12
WM-WF1	.03	-.04	.04	.11
BM-WM	.56**	-.12	.22	-.08
BF-WF1	.40**	-.00	.07	-.08
BM-WF1	.50**	-.08	.17	-.06
BF-WM	.30*	-.01	-.14	-.12
WF1-WF2	.37**	.11	.06	-.09

Note: All correlations are based on 72 items.

* $p < .02$
** $p < .002$

Table 3 contains Spearman rank-order correlation coefficients between BICAL-based measures of bias (and invariance) and various Rasch tests of fit provided by the BICAL program. The test-of-fit measures, which are averaged over the five groups of examinees, include the absolute value of the overall t test of fit, the between groups t test of fit, the absolute value of the deviation from unity of the discrimination index, and the standard error of estimated item difficulty.

Correlations between bias and Rasch tests of fit range from -.13 to .30. Although two correlation coefficients are marginally significant (.28 and .30, $p < .02$), there appears to be very little systematic relationship between bias and any of the tests of fit. An examination of the correlations between measures of invariance and the tests of fit in Table 3 reveals a more interesting pattern. Each of the five examinee group comparisons, that demonstrated significant correlations between bias and item position (see Table 2) and the greatest amount of differential test speededness (see Table 1), has a significant correlation between the invariance measure and the (between groups) t test fit or (the absolute value of the deviation from unity of) the discrimination index or both. The correlations range from .24 to .49 for the five group comparisons and the two tests of fit. This result is reasonable since both of these tests of fit are indicators of item fit across the range of number-right scores obtained within each examinee group.

The following results were obtained from the simulation study which was previously described. The 500 simulated ability parameters from the Gaussian normal distribution have an observed mean of .06 and an observed standard deviation of .92. The simulated ability parameters were used to generate estimated true scores for the first and second halves of the 72-item vocabulary test, by assuming that the item difficulty estimates based on the five examinee groups are the true values for each group. The differences in mean estimated true scores between half tests are presented in Table 4. Notice that the relationship between item difficulty estimates and item position, noted previously for BICAL estimates, results in considerably different mean estimated true scores for the two half tests. The lack of relationship between item difficulty estimates and item position for the LOGIST estimates is also reflected in Table 4. Also note that the relative magnitudes of the differences in half-test means based on the BICAL estimates are consistent with the speededness information contained in Table 1.

A caveat is in order concerning the information provided in Table 4. The values in the table should be considered as illustrative rather than definitive. The particular values of the differences depend on the characteristics of the simulated ability distribution. A convenient, relatively plausible, distribution was chosen to demonstrate the effect of test speededness upon BICAL and LOGIST item difficulty estimates. The application of a common set of simulated abilities to item difficulty estimates derived from the various examinee groups was designed to give the reader a feel for the relative impact of test speededness resulting from the use of the BICAL versus the LOGIST program.

Table 3

Spearman Rank-Order Correlations Between BICAL-Based Measures of Item Bias and Invariance

Group Comparison	Bias				Lack of Invariance			
	Overall T	Between T	Disc.	S.E.	Overall T	Between T	Disc.	S.E.
BM-BF	.27	.20	.30*	.08	-.12	-.09	-.09	.19
WM-WF1	.10	.11	.11	.14	.08	.06	.09	-.00
BM-WM	.24	.02	.11	.17	.25	.32*	.25	.03
BF-WF1	.15	-.08	.04	.28*	.20	.41**	.49**	.08
BM-WF1	.27	.05	.16	.22	.25	.40**	.29*	.05
BF-WM	.06	-.13	-.08	.10	.16	.30*	.36**	.08
WF1-WF2	.23	.04	.20	.23	.09	.24	.29*	.26

Note: All correlations are based on 72 items.

* $p < .02$
** $p < .002$

Table 4

Differences Between Mean True-Score Estimates for the First and Second Halves of the Vocabulary Test

Group upon which Estimates are Based	Difference	
	BICAL	LOGIST
BM	4.29*	-.35
BF	3.40*	-.01
WM	1.85*	.16
WF1	1.79*	.31
WF2	.74	-.03

* $p < .001$

Discussion

Most item response models, including the one-parameter logistic, assume unidimensionality, i.e., that a single person-ability parameter determines the probability that an examinee will respond correctly to an item. A speeded test clearly violates this assumption, since at least two abilities determine the probability of an examinee answering an item correctly, knowledge of the material covered by the test and the speed at which he/she moves through the test. For the examinee who fails to finish an examination, the earlier items tend to measure knowledge of the material tested and the later items tend to measure quickness in answering the questions. The LOGIST computer program is designed to compensate for the effects of speededness by maximizing a criterion function that does not involve items not reached by an examinee. The BICAL program in effect predicts that all items not reached by an examinee would have been incorrect responses. Poor results from the BICAL program for a speeded test are entirely predictable. The extent to which the modifications to the likelihood function made in the LOGIST program appear to be successful, even for a differentially speeded test, is noteworthy. The LOGIST program does seem to successfully reduce speed as a source of multidimensionality. This is particulary useful when item response models are being applied to existing tests or testing systems. The option of changing an existing test to make it essentially unspeeded is often not feasible or may not be desirable. On both conceptual and empirical grounds, LOGIST appears to be a much better choice than BICAL for parameter estimation involving partially speeded tests.

For those interested in using the BICAL computer program and the associated tests of fit, the results are mixed. The lack of any systematic relationship between item bias and the Rasch tests of fit implies that the tests of fit may not be useful in detecting biased items. However, the relationship between measures of item difficulty invariance and the between group t test and the absolute value of the deviation of the discrimination index from unity provides a more hopeful sign. Both of these indices are related to discrepancies in item difficulty estimates across examinee groups. This relationship was not found for the total t test or the standard error of the item difficulties. It should be emphasized that these results are based on only one test and on only one obvious source of multidimensionality, test speededness.

The results of the study suggest the following three recommendations:

1) The LOGIST computer program, rather than the BICAL program, should be used for partially speeded tests.

2) When using the BICAL computer program, give the between groups t test and the discrimination index more weight than the overall t test or the standard error of the item difficulties as indicators of poor item fit.

3) Do not expect Rasch tests of fit to aid in identification of biased items.

The results of the study emphasize two general points. First, the choice of an estimation procedure can be as crucial as the choice of a measurement model. Second, if the assumptions of a model are not met, the desirable properties of the model may or may not be obtained. Since the assumptions of a model are never entirely met in the practical world of testing, the problem becomes one of determining to what extent and in what ways model assumptions can be violated without losing the desirable features of the model.

REFERENCES

Angoff, W. H. Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971, pp. 508-600.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968

Douglass, J. B. A comparison of item response theory models for a classroom examination system. Unpublished doctoral dissertation, Michigan State University, 1980.

Hambleton, R. K., & Cook, L. L. Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 1977, 14, 75-96.

Hambleton, R. K., Swaminathan, H., Cook, L., Eignor, D., & Gifford, J. Developments in latent trait theory: models, technical issues and applications. Review of Educational Research, 1978, 48, 467-510.

Lord, F. M. An empirical study of item-test regression. Psychometrika, 1965, 30, 373-376.

Lord, F. M. An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. Educational and Psychological Measurement, 1968, 28, 989-1020.

Lord, F. M. Estimation of latent ability and item parameters when there are omitted responses. Psychometrika, 1974, 39, 247-264.

Lord, F. M. The 'ability' scale in item characteristic curve theory. Psychometrika, 1975, 44, 205-217.(a)

Lord, F. M. Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters (ETS RB 75-33). Princeton, N.J.: Educational Testing Service, 1975.(b)

Lord, F. M. Practical applications of item characteristic curve theory. Journal of Educational Measurement, 1977, 14, 117-138.

Lord, F. M. Applications of item response theory to practical testing problems. Hillsdale, N.J.: Erlbaum, 1980.

Lord, F. M. & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Marco, G. L. Item characteristic curve solutions to three intractable testing problems. Journal of Educational Measurement, 1977, 14, 139-160.

Rasch, G. Probabilitistic models for some intelligence and attainment tests. Copenhagen, Denmark: Denmarks Paedogogiske Institute, 1960.

Reckase, M. D. Unifactor latent trait models applied to multifactor tests: results and implications. Journal of Educational Statistics, 1979, 4, 207-230.

Ree, J. M. Estimating item characteristic curves. Applied Psychological Measurement, 1979, 3, 371-385.

Rentz, R. R. & Beshaw, W. L. Equating reading tests with the Rasch model, Volume I final report, Volume II technical reference tables. Athens, Ga.: University of Georgia, Educational Research Laboratory, 1975.

Rudner, L. M., Getson, P. R., and Knight, D. L. Biased item detection techniques. Journal of Educational Statistics, 1980, 5, 213-233.

Wood, R. L., Wingersky, M. S., & Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters. (Research Memorandum 76-6). Princeton, N.J.: Educational Testing Service, 1976.

Wright, B. D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

Wright, B. D., Mead, R., & Bell, S. BICAL: calibrating items with the Rasch model. (Research Memorandum No. 23B). Chicago: University of Chicago, Statistical Laboratory, Department of Education, 1979.

Wright, B. D., & Stone, M. Best test design: Rasch measurement. Chicago, Ill.: MESA, 1979.